Journal of Athletic Training 2012;47(3):297–305
doi: 10.4085/1062-6050-47.3.13

www.nata.org/journal-of-athletic-training

original research

Age-Related Differences and Reliability on Computerized and Paper-and-Pencil Neurocognitive Assessment Batteries

Johna K. Register-Mihalik, PhD, ATC*†; Daniel L. Kontos, MA, ATC‡; Kevin M. Guskiewicz, PhD, ATC, FNATA, FACSM†§; Jason P. Mihalik, PhD, CAT(C), ATC†§; Robert Conder, PsyD‖; Edgar W. Shields, PhD‡

*Clinical Research Unit, Emergency Services Institute, WakeMed Health & Hospitals and †Matthew Gfeller Sport-Related Traumatic Brain Injury Research Center, Raleigh, NC; ‡Department of Exercise and Sport Science and §Curriculum in Human Movement Science, Department of Allied Health Sciences, The University of North Carolina at Chapel Hill; ‖Carolina Neuropsychological Service, Raleigh, NC

Context: Neurocognitive testing is a recommended component in a concussion assessment. Clinicians should be aware of age and practice effects on these measures to ensure appropriate understanding of results.

Objective: To assess age and practice effects on computerized and paper-and-pencil neurocognitive testing batteries in collegiate and high school athletes.

Design: Cohort study.

Setting: Classroom and laboratory.

Patients or Other Participants: Participants consisted of 20 collegiate student-athletes (age = 20.00 ± 0.79 years) and 20 high school student-athletes (age = 16.00 ± 0.86 years).

Main Outcome Measure(s): Hopkins Verbal Learning Test scores, Brief Visual-Spatial Memory Test scores, Trail Making Test B total time, Symbol Digit Modalities Test score, Stroop Test total score, and 5 composite scores from the Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT) served as outcome measures. Mixed-model analyses of variance were used to examine each measure.

Results: Collegiate student-athletes performed better than high school student-athletes on ImPACT processing speed composite score ($F_{1,38} = 5.03$, $P = .031$) at all time points. No other age effects were observed. The Trail Making Test B total time ($F_{2,66} = 73.432$, $P < .001$), Stroop Test total score ($F_{2,76} = 96.85$, $P = < .001$) and ImPACT processing speed composite score ($F_{2,76} = 5.81$, $P = .005$) improved in test sessions 2 and 3 compared with test session 1. Intraclass correlation coefficient calculations demonstrated values ranging from 0.12 to 0.72.

Conclusions: An athlete's neurocognitive performance may vary across sessions. It is important for clinicians to know the reliability and precision of these tests in order to properly interpret test scores.

Key Words: concussions, traumatic brain injuries, serial testing

Key Points

- An athlete's neurocognitive test performance may vary across serial testing sessions. To properly interpret score variations, the clinician must know the reliability and precision of these tests.
- With practice effects, the greatest improvement in test scores occurs between the first and second administrations of a neurocognitive test.
- Baseline measures of processing speed may need to be reassessed as an athlete ages.

Concussion is a serious injury that occurs at all levels of sport and can affect the cognitive, physical, and behavioral abilities of an athlete.[1–8] Therefore, it is important that these injuries be properly evaluated and managed. Currently, a comprehensive evaluation is recommended, in which clinical attributes, symptoms, neurocognitive performance, and balance are assessed.[9–11] This multifaceted clinical model is more than 90% sensitive in identifying concussion;[12] however, when any of these measures is used in isolation, the sensitivity often drops to less than 60%.[12] One important component of this evaluation is the neurocognitive assessment. This can be performed using either traditional paper-and-pencil neurocognitive tests or computerized neurocognitive tests. For reasons including ease of administration, reduced testing time, and availability of the tests to sports medicine clinicians, computerized neurocognitive tests have gained considerable popularity in sports medicine settings. Despite their widespread use, the psychometric properties, practice effects, and age effects of neurocognitive testing are not well understood in the athletic population.

The vast majority of people participating in contact and collision sports are under 19 years of age,[13] and an earlier study[14] suggested that high school athletes may be more at risk for concussion than college athletes. Given the number of athletes participating across many age levels, clinicians should be mindful of any age effects in the interpretation of both baseline and postinjury data. Furthermore, high

school athletes may recover at a slower rate than do collegiate athletes after a concussion.[15]

Performance on many neurocognitive tests may be improved by prior exposure to testing stimuli and procedures[16] in the absence of any actual recovery by the patient.[16] This false improvement is due to 2 factors: the athlete has already learned the procedures involved in taking the test, and he or she already knows the specific content of the tests.[17] Improvement in test performance due to practice effects may cause inflated neurocognitive test scores, which can mimic neurocognitive recovery, and may lead to returning an athlete to competition prematurely. An example of these practice effects occurs with processing speed on the Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT).[18] Conversely, lack of improvement with serial assessments on neurocognitive tests can imply continued concussive impairment.[19]

Serial (repeated) testing is often used to track an athlete's neurocognitive recovery over time.[2,4,19] The consistency of an individual's performance on these measures must be carefully considered in the context of interpreting post-injury serial neurocognitive testing. Reliability relates to consistency of measurement on the same task at different time intervals.[20,21] Reliability is a property of tests that can be established by various statistical methods: test-retest, alternate forms, or split-half or internal consistency. It is examined and estimated through the systematic and ongoing evaluation of different kinds of reliability evidence, applied in different clinical and nonclinical contexts, to numerous groups. Efforts have been made to address these psychometric issues in recent studies.[19,22,23] One recent investigation[22] suggested that 3 of the most commonly used computerized tests had only low to moderate test-retest reliability across test sessions[22]; however, the number of days between test sessions constituted a wide range. In addition, participants in this study were asked to complete 3 computerized testing batteries during each of the 3 testing sessions included in the study, which may not be clinically applicable.

These psychometric factors should not be dismissed. Rather, they should be viewed as practical and essential considerations when using cognitive testing to make concussion management and return-to-play decisions. Therefore, the purpose of our study was to examine age-group differences (collegiate versus high school) on neurocognitive performance in healthy athletes with no history of concussion in the last 5 years. Two additional aims were to compare the consistency (reliability) of responses across all participants and practice effects between the age groups.

METHODS

Participants

Forty healthy and active volunteers participated in this study. Participants consisted of 20 National Collegiate Athletic Association Division I student-athletes (age = 20.00 ± 0.79 years) and 20 student-athletes (age = 16.00 ± 0.86 years) from 2 high schools. Each age group contained 10 males and 10 females. Participants were classified as *healthy* if they had no history of diagnosed concussion within the last 5 years and no known neurologic, psychiatric, or psychological conditions that would affect cognition. They were classified as *active athletes* if they engaged in athletics 3 or more days per week. Individuals 18 years of age were excluded in order to maintain a clear separation between the high school and collegiate groups. Effort level for each participant's scores was evaluated by assessing the ImPACT impulse control composite score: a score greater than 30 constituted an invalid test. All participants included in the study met the criteria for valid tests.

Instrumentation

Participants were tested on both a computer-based test battery and a traditional paper-and-pencil–based test battery to assess neurocognitive performance during 3 test sessions. Outcome measures for each test are presented in Tables 1 and 2. The paper-and-pencil battery was designed to include tests theoretically measuring cognitive domains similar to those assessed in the computerized battery.

The computer-based test battery used was the ImPACT (version 3; ImPACT Applications, Inc, Pittsburgh, PA). This computerized neurocognitive testing program assesses a number of cognitive processes, including visual and verbal memory, attention, working memory, processing speed, reaction time, impulse control, and response inhibition. The 6 subtests are Word Memory, Design Memory, X's and O's, Symbol Match, Color Match, and Three Letters. Five composite scores are provided in the clinical report: verbal memory, visual memory, processing speed, reaction time, and impulse control. A self-reported postconcussion symptom scale (PCSS) is also included in the ImPACT program. However, we did not analyze symptom scores as part of this study. Reliability of the ImPACT composite scores has been demonstrated with intraclass correlation coefficient (ICC) values ranging from 0.23 to 0.46 for verbal memory, 0.32 to 0.65 for visual memory, 0.38 to 0.75 for processing speed, 0.39 to 0.68 for reaction time, and 0.15 to 0.54 for impulse control.[22,24]

Paper-and-Pencil Battery

Hopkins Verbal Learning Test–Revised (HVLT-R).[25] This test is a measure of verbal learning and memory in which the clinician reads 12 words aloud to the athlete. The athlete then attempts to immediately free recall as many of the words as possible in any order. This process is repeated 2 more times for a total of 3 free-recall trials. The other paper-and-pencil tests described below are then completed. At the end of the traditional neurocognitive test battery, the athlete is asked to complete a delayed trial in which he or she tries to free recall as many words as possible from the original list. Lastly, a discrimination trial is completed, in which the clinician reads 24 words aloud to the athlete; 12 are from the original list, for which a response of *yes* is expected, and 12 additional words (6 semantically related false-positives and 6 semantically unrelated false-positives), for which a response of *no* is expected. Three alternate forms were used (A, B, and C) to reduce learning effects across testing sessions. Reliability for the HVLT-R is lower than for some of the other measures included in the study and ranges from 0.36 to 0.49.[26]

Brief Visuospatial Memory Test–Revised (BVMT-R).[27] This test assesses the participant's visual-spatial memory with 3 learning trials and a delayed free-recall trial.

Table 1. Main and Interaction Effects for Paper-and-Pencil Neuropsychological Test Scores

Test	Score	Age Group and Session Average	Session, Mean ± SD			Group Average, Mean ± SD	Main Effect		Interaction Effect
			1	2	3		Time[a]	Age[b]	
Hopkins Verbal Learning	Total recalled (immediate)	High school	26.30 ± 3.80	28.20 ± 3.25	27.75 ± 3.51	27.42 ± 3.56	$F_{2,76} = 2.68$, $P = .075$	$F_{1,38} = 0.04$, $P = .844$	$F_{2,76} = 1.81$, $P = .170$
		College	27.20 ± 3.49	27.60 ± 2.78	26.90 ± 3.71	27.23 ± 3.31			
		Session average	26.75 ± 3.63	27.90 ± 3.00	27.33 ± 3.59				
	Discrimination (immediate)	High school	11.75 ± 0.44	11.80 ± 0.41	11.85 ± 0.37	11.80 ± 0.40	$F_{2,76} = 0.63$, $P = .534$	$F_{1,38} = 2.00$, $P = .165$	$F_{2,76} = 0.34$, $P = .712$
		College	11.50 ± 0.76	11.65 ± 0.59	11.55 ± 0.94	11.57 ± 0.77			
		Session average	11.63 ± 0.63	11.73 ± 0.51	11.70 ± 0.72				
	Total recalled (delayed)	High school	9.80 ± 1.85	9.85 ± 2.11	9.90 ± 2.61	9.85 ± 2.18	$F_{2,76} = 0.03$, $P = .969$	$F_{1,38} = 0.73$, $P = .400$	$F_{2,76} = 0.14$, $P = .871$
		College	10.40 ± 2.14	10.40 ± 1.73	10.20 ± 2.12	10.33 ± 1.97			
		Session average	10.10 ± 2.00	10.13 ± 1.92	10.05 ± 2.35				
	Discrimination (delayed)	High school	11.60 ± 0.60	11.15 ± 1.04	11.55 ± 0.60	11.43 ± 0.79	$F_{2,65} = 1.37$, $P = .260$	$F_{1,38} = 0.22$, $P = .640$	$F_{2,65} = 0.61$, $P = .525$
		College	11.30 ± 1.30	11.25 ± 1.29	11.40 ± 1.10	11.32 ± 1.21			
		Session average	11.45 ± 1.01	11.20 ± 1.16	11.48 ± 0.88				
Brief Visuospatial Memory-Revised	Total recalled (immediate)	High school	32.75 ± 2.79	33.45 ± 1.47	32.65 ± 2.30	32.95 ± 2.24	$F_{2,76} = 3.19$, $P = .046$	$F_{1,38} = 0.35$, $P = .557$	$F_{2,76} = 0.29$, $P = .743$
		College	32.25 ± 3.99	33.35 ± 2.78	31.80 ± 4.47	32.47 ± 3.80			
		Session average	32.50 ± 3.40	33.40 ± 2.19	32.23 ± 3.53				
	Total recalled (delayed)	High school	11.85 ± 0.49	11.90 ± 0.31	11.65 ± 0.75	11.80 ± 0.55	$F_{2,64} = 3.36$, $P = .049$	$F_{1,38} = 0.71$, $P = .405$	$F_{2,64} = 0.17$, $P = .807$
		College	11.75 ± 0.79	11.85 ± 0.37	11.45 ± 0.94	11.68 ± 0.75			
		Session average	11.80 ± 0.65	11.88 ± 0.33	11.55 ± 0.85				
Trail Making Form B	Total time	High school	61.43 ± 13.79	47.36 ± 9.49	40.20 ± 11.28	49.66 ± 14.50	$F_{2,66} = 73.43$, $P < .001$[c,d]	$F_{1,38} = 6.16$, $P = .018$[e]	$F_{2.66} = 1.56$, $P = .216$
		College	50.04 ± 14.53	40.61 ± 12.13	33.74 ± 0.22	41.46 ± 13.71			
		Session average	55.73 ± 15.12	43.98 ± 11.28	36.97 ± 10.68				
Symbol Digit Modalities	Total	High school	41.00 ± 5.85	40.45 ± 6.25	41.95 ± 5.94	41.13 ± 5.95	$F_{2,76} = 0.91$, $P = .407$	$F_{1,38} = 0.55$, $P = .463$	$F_{2,76} = 0.24$, $P = .790$
		College	42.60 ± 6.55	42.15 ± 7.69	42.70 ± 5.89	42.48 ± 6.64			
		Session average	41.80 ± 6.18	41.30 ± 6.97	42.33 ± 5.85				
Stroop	Total	High school	52.95 ± 10.86	59.95 ± 11.28	63.95 ± 12.23	58.95 ± 12.17	$F_{2,76} = 96.85$, $P < .001$[c,d]	$F_{1,38} = 0.03$, $P = .857$	$F_{2,76} = 0.31$, $P = .737$
		College	52.90 ± 7.90	61.20 ± 10.50	64.55 ± 11.82	59.55 ± 11.18			
		Session average	52.93 ± 9.37	60.58 ± 10.77	64.25 ± 11.87				

[a] *F* and *P* values are associated with session means.
[b] *F* and *P* values are associated with high school versus college means.
[c] Main effect of time: session 2 performance was superior to session 1 performance.
[d] Main effect of time: session 3 performance was superior to session 1 performance.
[e] Main effect of group: performance of collegiate athletes was superior to that of high school athletes.

Table 2. Main Effects and Interaction Effects for the ImPACT Composite Scores

Test	Sample	Session, Mean ± SD			Group Average, Mean ± SD	Main Effect of Time[a]		Main Effect of Age[b]		Interaction Effect	
		1	2	3		$F_{2,76}$	P	$F_{1,38}$	P	$F_{2,76}$	P
Verbal memory	High school	89.20 ± 7.73	86.65 ± 8.44	89.10 ± 7.67	88.32 ± 7.91	0.37	0.69	0.33	0.569	1.28	0.284
	College	90.05 ± 6.48	90.30 ± 7.35	87.85 ± 10.79	89.40 ± 8.34						
	Total	89.63 ± 7.05	88.48 ± 8.03	88.48 ± 9.26							
Visual memory	High school	78.40 ± 9.34	79.95 ± 10.45	80.30 ± 7.46	79.55 ± 9.05	1.63	0.203	0.79	0.378	0.19	0.829
	College	79.50 ± 10.45	82.95 ± 9.73	83.00 ± 9.67	81.82 ± 10.86						
	Total	78.95 ± 11.20	81.45 ± 10.08	81.65 ± 8.63							
Processing speed	High school	39.43 ± 7.88	43.03 ± 7.36	43.58 ± 6.45	42.01 ± 7.34	5.81	.005[c,d]	5.03	.031[e]	2.23	0.114
	College	45.81 ± 6.03	46.86 ± 7.09	46.65 ± 6.55	46.44 ± 6.48						
	Total	42.62 ± 7.64	44.95 ± 7.39	45.11 ± 6.60							
Reaction time	High school	0.55 ± 0.06	0.52 ± 0.06	0.53 ± 0.08	0.53 ± 0.07	2.01	0.141	1.21	0.279	0.12	0.889
	College	0.52 ± 0.06	0.51 ± 0.08	0.51 ± 0.07	0.51 ± 0.07						
	Total	0.53 ± 0.06	0.52 ± 0.07	0.52 ± 0.08							
Impulse control	High school	9.20 ± 5.40	8.85 ± 7.51	8.75 ± 4.51	8.93 ± 5.84	0.09	0.961	2.53	0.12	0.39	0.678
	College	6.20 ± 4.31	6.65 ± 4.13	7.05 ± 5.29	6.63 ± 4.54						
	Total	7.70 ± 5.05	7.75 ± 6.09	7.90 ± 4.92							

[a] F and P values are associated with session means.
[b] F and P values are associated with high school versus college means.
[c] Main effect of time: session 2 performance was superior to session 1 performance.
[d] Main effect of time: session 3 performance was superior to session 1 performance.
[e] Main effect of group: performance of collegiate athletes was superior to that of high school athletes.

Participants must learn and reproduce 6 abstract designs arranged in 2 columns and 3 rows. Three alternate forms were used (1, 2, and 3) to reduce learning effects across testing sessions. The BVMT-R is moderately to highly reliable, with values ranging from 0.73 to 0.91.[28]

Trail Making Test Form B (TMT-B; Trails B).[29] This test assesses the participant's visual scanning, attention, mental flexibility, and visual-motor speed. The TMT-B requires the participant to draw a continuous line connecting circles in ascending order, alternating between number (1 through 13) and letter (A through K). The score is the time, in seconds, required to complete the test. Only 1 form was used for this test. Because of previous findings regarding sport-related concussion,[2] we omitted the Trail Making Test Form A (Trails A) from our testing battery and only included TMT-B. In most previously published traditional neuropsychology literature, Trails A is completed before Trails B; however, Trails B is more sensitive than Trails A for cognitive flexibility,[30] brain dysfunction, and sport concussions.[2] Other authors[2] have used Trails B exclusively. Reliability of the TMT-B is moderate to high, ranging from 0.65 to 0.85.[31,32]

Symbol Digit Modalities Test (SDMT).[33] This test assesses psychomotor speed, visual short-term memory, attention, and concentration. Participants are asked to fill in a series of empty boxes underneath symbols with the corresponding number, using a key on top of the test form to identify which number goes with each symbol. The score was calculated as the number of correct responses in 60 seconds (abbreviated from the typical 90 seconds to shorten overall testing time). Three alternate forms were used (A, B, and C) to reduce learning effects across test sessions. Reliability ranges from 0.82 to 0.87.[32,34]

Stroop Test.[35] This test assesses speed of processing and cognitive flexibility. A participant is given a page with columns of color names (red, green, and blue), which are printed in different font colors, and asked to say the name of the font color in which each word is printed, ignoring the word that was spelled out. For example, if the printed word "red" appears in blue color font, the answer should be "blue." The person is given 45 seconds to correctly identify the font color of as many words as possible. We did not administer the word reading and color naming subtests. Only the color-word subtest was given, and only 1 form was used for this test. Reliability ranges from 0.54 to 0.60.[36,37]

Procedures

High school participants reported to a classroom at their respective high schools, and collegiate participants reported to a sports medicine research laboratory. A single certified athletic trainer, trained in the administration of neurocognitive testing, administered all tests. No more than 2 participants were tested at the same time. All participants reported to their respective testing site for a total of 3 visits, with at least 24 hours but no more than 72 hours between visits (average time between sessions 1 and 2 = 1.8 ± 0.61 days and between sessions 2 and 3 = 1.6 ± 0.59 days). Each testing session lasted for approximately 1 hour. All participants completed both the ImPACT and the traditional battery of 5 paper-and-pencil tests in counterbalanced order for each test session. The first participant determined which test battery (computerized or traditional) to begin by random selection (ie, coin flip). All of the following participants began with the test battery that counterbalanced the previous participant and used the same test battery order for all 3 test sessions.

The rate of testing (how fast the individual completed the test) was determined by the participant for both the ImPACT and paper-and-pencil test batteries. Upon completion of 1 test, the participant confirmed that he or she was ready to proceed to the next test and continued until all tests for that battery were completed. Upon completion of 1 test battery, the participant began the remaining test battery after a 5-minute rest period. The test session was concluded once both test batteries were completed.

Table 3. Effect Sizes for Outcome Measures

Test	Effect Size, μ_p^2		
	Interaction	Time	Age
Hopkins Verbal Learning-Revised			
Total recalled (immediate)	0.046	0.066	0.001
Discrimination index	0.009	0.016	0.05
Total recalled (delayed)	0.004	0.001	0.019
Discrimination index (delayed)	0.016	0.035	0.006
Brief Visuospatial Memory-Revised			
Total recalled (immediate)	0.008	0.078	0.009
Total recalled (delayed)	0.004	0.081	0.018
Trail Making Form B	0.039	0.659	0.14
Symbol Digit Modalities total score	0.006	0.023	0.014
Stroop total score	0.008	0.718	0.001
ImPACT verbal memory	0.033	0.01	0.009
ImPACT visual memory	0.005	0.041	0.02
ImPACT processing speed	0.056	0.133	0.117
ImPACT reaction time	0.003	0.05	0.031
ImPACT impulse control	0.01	0.001	0.062

Data Analysis

One 2 × 3 mixed-model analysis of variance (age × time) was calculated for each of the 14 outcome measures. For each outcome measure, analysis of variance was conducted to examine the main effects for group (age) and test time (practice effects) to determine differences between collegiate and high school athletes for the ImPACT and paper-and-pencil neurocognitive test scores. Interaction effects were analyzed to examine the joint effects of age and test time (practice) for each outcome measure.

An ICC [2,1] with standard error of measurement (SEM) was calculated to determine the consistency of the athletes' performance across serial neurocognitive tests for each of the 14 clinically relevant outcome measures. Pearson bivariate correlations were used to examine correlation of these measures across time in the combined sample. The change scores within each group (college and high school) were compared with the previously published reliable change indices (RCIs) for the ImPACT composite measures.[19] This comparison was made to the table given in the article[19] that provided established RCIs for each of the composite measures produced by ImPACT, allowing both researchers and clinicians to see if the change occurring across test sessions is a meaningful change.

To analyze the data, we used SPSS (version 16.0; SPSS Inc, Chicago, IL). Mean scores and standard deviations were calculated for each outcome measure. An a priori α level of significance was set at .05 for all analyses. Because our 9 paper-and-pencil outcome measures may be related, we adjusted our level of significance to .0056 for all analyses related to the paper-and-pencil testing battery. The .05 level was applied for the ImPACT composite score measures. We calculated that 20 participants per group would be needed for an effect size of 0.80 and power of 0.80.

RESULTS

Effects of Age and Practice

No significant interaction effects were noted for the computerized or paper-and-pencil batteries (Tables 1 and 2). No statistical differences were observed for the effect of age on any of the paper-and-pencil outcome measures. A main effect of age was observed for ImPACT processing speed score ($F_{1,38} = 5.03$, $P = .031$) whereby college students performed better than did high school students. Effect sizes for the main effects and interaction effect related to each outcome measure are shown in Table 3.

A main effect of test session was seen for TMT-B total time ($F_{2,66} = 73.43$, $P < .001$), Stroop Test total score ($F_{2,76} = 96.85$, $P < .001$), and ImPACT processing speed composite score ($F_{2,76} = 5.81$, $P = .005$). For each of these measures, averages for test sessions 2 and 3 were significantly better than for session 1 (Tables 1 and 2), with improvement of 22% from sessions 1 and 2 and improvement of more than 30% from session 1 to session 3 for TMT-B. The ImPACT processing speed scores improved by more than 5% from session 1 to session 2 and from session 1 to session 3. Performance on the Stroop Test

Table 4. Reliability and Precision of Paper-and-Pencil Neuropsychological Test Scores

Test	Intraclass Correlation Coefficient [2,1]	Standard Error of Measurement	Test-Retest Correlations					
			r_{12}[a]	P_{12}[a]	r_{23}[b]	P_{23}[b]	r_{13}[c]	P_{13}[c]
Hopkins Verbal Learning-Revised								
Total recalled (immediate)	0.56	2.412	0.485	.002	0.681	<.001	0.561	<.001
Discrimination index (immediate)	0.57	0.471	0.474	.002	0.61	<.001	0.649	<.001
Total recalled (delayed)	0.59	1.498	0.51	.001	0.616	<.001	0.643	<.001
Discrimination index (delayed)	0.3	0.97	0.206	.203	0.535	<.001	0.389	.013
Brief Visuospatial Memory-Revised								
Total recalled (immediate)	0.5	2.485	0.361	.022	0.508	.001	0.694	<.001
Total recalled (delayed)	0.12	0.799	0.354	.025	−0.023	.89	0.299	.061
Trail Making Form B								
Total time	0.39	11.8	0.668	<.001	0.755	<.001	0.719	<.001
Symbol Digit Modalities Total score	0.72	3.691	0.795	<.001	0.743	<.001	0.621	<.001
Stroop								
Total score	0.69	6.659	0.899	<.001	0.918	<.001	0.864	<.001

[a] Indicates time 1 to time 2.
[b] Indicates time 2 to time 3.
[c] Indicates time 1 to time 3.

Table 5. Reliability and Precision of ImPACT Composite Scores

ImPACT Composite Score	Intraclass Correlation Coefficient [2,1]	Standard Error of Measurement	Test-Retest Correlations					
			r_{12}[a]	P_{12}[a]	r_{23}[b]	P_{23}[b]	r_{13}[c]	P_{13}[c]
Verbal memory	0.29	7.809	0.192	.235	0.273	.088	0.396	.013
Visual memory	0.45	8.27	0.547	<.001	0.483	.002	0.358	.023
Processing speed	0.71	4.094	0.828	<.001	0.723	<.001	0.654	<.001
Reaction time	0.6	0.051	0.757	<.001	0.659	<.001	0.629	<.001
Impulse control	0.63	3.699	0.618	<.001	0.646	<.001	0.641	<.001

[a] Indicates time 1 to time 2.
[b] Indicates time 2 to time 3.
[c] Indicates time 1 to time 3.

improved by 14% from session 1 to session 2 and by 21% from session 1 and session 3. No differences were seen between sessions 2 and 3, suggesting that test scores had stabilized by this point. Effect sizes for all measures are presented in Table 3.

Reliability and Precision

The ICC values ranged from 0.12 to 0.72. The 3 lowest values were for BVMT-R total delayed recall (ICC [2,1] = 0.12), ImPACT verbal memory composite score (ICC [2,1] = 0.29), and HVLT-R delayed discrimination index (ICC [2,1] = 0.30). The 3 highest values were for SDMT total score (ICC [2,1] = 0.72), ImPACT processing speed composite score (ICC [2,1] = 0.71), and Stroop total score (ICC [2,1] = 0.69). Test-retest correlations ranged from low to high depending on the outcome. Lists of ICCs, SEMs, and test-retest correlations for all variables are provided in Tables 4 and 5. Table 6 includes comparisons of our data with previously published reliable change indices[19] and represents the percentage of athletes within each group whose performance across test sessions changed reliably. We included this table to illustrate the percentage of participants in the sample whose scores changed at clinically meaningful levels.

DISCUSSION

Overall, our most clinically relevant findings concern the variability in performance on measures included in both the computerized and paper-and-pencil testing across sessions. These findings were most noticeable between test sessions 1 and 2. These results highlight the need to understand this variability and to control for as many factors as possible to produce more stable results across serial testing sessions.

Age and Practice Effects

One purpose of our study was to determine if age affects neurocognitive test performance. Of all the tests we used, age-related differences were found for only ImPACT processing speed composite scores, with collegiate athletes performing better than high school athletes across all 3 test sessions. This result adds support to the finding of Iverson et al[38] that adolescents (ages 13–18 years) displayed age effects for processing speed. Hunt and Ferrara[39] observed age-related differences among high school students on the TMT-B and suggested that measures of processing speed may differ by age group. Presumably these differences reflect ongoing brain development between adolescence and early adulthood[40,41] and may reflect underlying neuromaturational processes. These results are consistent with the changes in cognitive maturity and decline described in the current literature.[40,41]

Clinicians should be aware of this difference between age groups when evaluating an athlete's performance on processing-speed measures. Most importantly, these findings suggest that a baseline score (at least on these types of measures) for a young athlete should be reassessed once he or she reaches college. Future researchers should continue to monitor the effects of age and study a larger spectrum of age groups, including athletes at the high school, college, and even professional levels.

In addition, practice effects were similar across the collegiate and high school athletes, with the most drastic improvements occurring between test sessions 1 and 2. Overall, both groups displayed improvement on both the immediate and delayed portions of the TMT-B, Stroop Test, and ImPACT processing speed composite score. These results reflect those in similar studies[19,23] and indicate that some orientation to a task may be needed to obtain a stable baseline measure. However, in a different patient

Table 6. Athletes with Reliable Change (Percent Improvement or Decline) Across Test Sessions on ImPACT Composite Scores

ImPACT Composite Score	80% Confidence Interval[a]	High School (n = 20)				College (n = 20)			
		Time 2 to Time 1	Time 3 to Time 1	Time 3 to Time 2	Overall[b]	Time 2 to Time 1	Time 3 to Time 1	Time 3 to Time 2	Overall[b]
Verbal memory	8.75	12 (60.0)	9 (45.0)	6 (30.0)	17 (85.0)	9 (45.0)	11 (55.0)	11 (55.0)	15 (85.0)
Visual memory	13.55	2 (10.0)	5 (25.0)	3 (15.0)	7 (35.0)	5 (25.0)	4 (20.0)	4 (20.0)	9 (45.0)
Reaction time	0.06	4 (20.0)	12 (60.0)	8 (40.0)	15 (75.0)	3 (15.0)	6 (30.0)	4 (20.0)	9 (45.0)
Processing speed	4.98	7 (35.0)	8 (40.0)	4 (20.0)	12 (60.0)	7 (35.0)	4 (20.0)	7 (35.0)	7 (35.0)

[a] Based on Iverson et al.[19]
[b] Percentage of athletes within each group who had a reliable change in score across at least 2 test sessions.

population, previous authors[42] suggested that dual baseline testing (obtaining a second baseline measurement) may be most useful to obtain a stabilized baseline score on some measures of cognitive function. In addition, this may indicate that individuals who have not taken a baseline test in a few months or years and undergo postinjury tests on 2 days in close proximity may exhibit practice effects between those sessions. Also, the lack of a learning effect after multiple postinjury test sessions may itself indicate deficits,[43] although few investigators have examined this point empirically. Furthermore, even though the previous literature has identified practice effects on many measures of neurocognitive function, a variety of factors may cause the variability in performance. Therefore, practice effects should often be interpreted with caution.[44] Regarding the influence of practice, the greatest effects were seen on the 2 paper-and-pencil neurocognitive tests for which there was only a single form each: TMT-B and Stroop Test. Significantly lower completion times for Trails B were noted both for the high school and collegiate groups (Trails B is scored by the amount of time to task completion; a lower time reflects better performance) in sessions after the initial test. Improved performance in correct color-word reading was noted on the Stroop Test. These results highlight the false improvement that can occur from an athlete's repeated exposure to the same form of test and result in premature return to play.

Among the pencil-and-paper tests with alternate (and presumably equivalent) forms, observed practice effects were minimal and statistically nonsignificant. This was shown especially with the lack of practice effects on the BVMT-R, HVLT-R, and SDMT. However, although alternative forms can factor out the content practice effect, they do not factor out the procedural practice effect from identical test instructions.[17] On the ImPACT, practice effects on processing speed composite were similar to those reported by Iverson et al in 2003.[19]

Reliability and Consistency

In sports medicine, neurocognitive testing is used to identify cognitive deficits and track an athlete's improvement over time. Consistency of the athlete's performance and stability of the actual measure are difficult to differentiate. Regardless, performance consistency during serial neurocognitive testing becomes critical for an accurate evaluation. The reliabilities across testing sessions within our sample ranged from low to moderate, indicating a need for further investigation into the stability and consistency of these measures over time. In a previous study[45] using an alternate computerized battery, an aggregated score from all outcomes on the Automated Neuropsychological Assessment Metrics displayed high reliability, but the individual module scores were similar to those we observed. Two previous groups[22,46] examined the reliability of ImPACT scores and found similar results; however, these authors assessed reliability over a longer timeframe.

A few observations may help to explain the range of reliability measures. The HVLT-R discrimination index (delayed) and the BVMT-R total recall (delayed) showed high ceiling effects for absolute scores, with little to no variability across test sessions. This lack of variability may have confounded the ICC results. Although no ceiling effect for scores was seen with TMT-B total time, ImPACT verbal memory composite score, or ImPACT visual memory composite score, reliability measures were low. Another factor that appeared to affect reliability was the time limit. In our study, tests with a set time limit for completion resulted in higher ICC values than did tests with no time limit. The SDMT, Stroop Test, and ImPACT processing speed composite had set time limits for completion and had the highest reliability values. This known endpoint of the test may increase motivation for the test taker, resulting in a more accurate representation of the individual's highest potential from one test to the next. To further examine stability over time, we compared our results using the reliable change methods proposed for the ImPACT[19] (Table 5). This method can account for test psychometric values and an individual's performance, which may be useful given the variability in these measures. The method has been suggested as a useful tool in understanding what represents clinical change after a concussive injury.[19,47,48] More than 35% of both our high school and collegiate athletes performed significantly better or worse across at least 2 sessions on all composite measures of the ImPACT. This indicates that performance may differ across testing time points, specifically when compared with the initial test session, and may reflect familiarity with the task and learning effects.

Test reliability and precision should be carefully considered by the clinician conducting serial neurocognitive testing in an athletic population. Based on ICC values and test-retest correlations, measures such as the HVLT-R total recalled (immediate and delayed), Stroop Test, and SDMT total score may be more appropriate for serial neurocognitive testing than the TMT-B total time, ImPACT verbal memory composite score, or ImPACT visual memory composite score. Variability is likely to occur across any serial neurocognitive tests, but these ICC and SEM values may give the clinician a better understanding of how much variability to expect from one test to the next. Future researchers should continue to explore the consistency of athletes' performance across serial neurocognitive tests. Increased duration of serial neurocognitive testing may provide a more accurate measure of each athlete's performance over time. In addition, alternative analyses that account for high ceiling effects may be ideal.

Limitations

As with any study, ours is not without limitations. A small window of time was allowed between test sessions; often the period of time to the initial postinjury session from baseline is longer. In addition, we only used a few of the paper-and-pencil tests available and 1 computerized test battery, which may limit our findings to these particular batteries. Lastly, the study had a relatively small sample size. However, given the effect sizes observed in the study, the lack of differences observed was most likely not clinically meaningful.

CONCLUSIONS

Outcomes of this study warrant attention from clinicians who are tasked with caring for athletes at risk of sport-related concussion. We demonstrated that athletes' neurocognitive test performances may vary across serial testing sessions. It is important for the clinician to know the reliability and

precision of these tests in order to properly interpret the variations in test scores. In some cases, the variability across serial neurocognitive testing is due to practice effects.

In the presence of a practice effect, the clinician can expect the greatest improvement in test scores to occur between the first and second administrations of a neurocognitive test. The clinician must be able to differentiate between a learning effect and neurocognitive recovery so as to make an accurate decision about whether the concussed athlete has recovered and is ready to return to competition. In addition, this finding of low to moderate reliability further illustrates the need for trained neuropsychologists to assist in the interpretation of neurocognitive testing results because many factors can influence the variability and accuracy of these scores.

This study also illustrated that for tests of processing speed, age-related differences exist between high school and collegiate athletes, with a much higher percentage of high school athletes showing improvements on reaction time and processing speed variables across test sessions. Therefore, at a minimum, baseline measures of processing speed may need to be reassessed as an athlete ages to ensure the most accurate representation of proper cognitive function due to continued brain development, among other factors. Accurate baseline assessments are important because depressed baseline levels may lead to faulty interpretation of postinjury results and possible premature return to play.

REFERENCES

1. Collins MW, Lovell MR, Iverson GL, Cantu RC, Maroon JC, Field M. Cumulative effects of concussion in high school athletes. *Neurosurgery*. 2002;51(5):1175–1181.
2. Guskiewicz KM, Ross SE, Marshall SW. Postural stability and neuropsychological deficits after concussion in collegiate athletes. *J Athl Train*. 2001;36(3):263–273.
3. Iverson GL, Brooks BL, Collins MW, Lovell MR. Tracking neuropsychological recovery following concussion in sport. *Brain Inj*. 2006;20(3):245–252.
4. Lovell MR, Collins MW, Iverson GL, et al. Recovery from mild concussion in high school athletes. *J Neurosurg*. 2003;98(2):296–301.
5. Lovell MR, Collins MW, Iverson GL, Johnston KM, Bradley JP. Grade 1 or ''ding'' concussions in high school athletes. *Am J Sports Med*. 2004;32(1):47–54.
6. Lincoln AE, Caswell SV, Almquist JL, Dunn RE, Norris JB, Hinton RY. Trends in concussion incidence in high school sports: a prospective 11-year study. *Am J Sports Med*. 2011;39(5):958–963.
7. Daneshvar DH, Nowinski CJ, McKee AC, Cantu RC. The epidemiology of sport-related concussion. *Clin Sports Med*. 2011;30(1):1–17, vii.
8. Bakhos LL, Lockhart GR, Myers R, Linakis JG. Emergency department visits for concussion in young child athletes. *Pediatrics*. 2010;126(3):e550–e556.
9. Guskiewicz KM, Bruce SL, Cantu RC, et al. National Athletic Trainers' Association position statement: management of sport-related concussion. *J Athl Train*. 2004;39(3):280–297.
10. McCrory P, Meeuwisse W, Johnston K, et al. Consensus statement on concussion in sport: The 3rd International Conference on Concussion in Sport held in Zurich, November 2008. *Br J Sports Med*. 2009;43(suppl 1):i76–i90.
11. McCrory P, Meeuwisse W, Johnston K, et al. Consensus statement on concussion in sport: 3rd International Conference on Concussion in Sport held in Zurich, November 2008. *Clin J Sport Med*. 2009;19(3):185–200.
12. Broglio SP, Macciocchi SN, Ferrara MS. Sensitivity of the concussion assessment battery. *Neurosurgery*. 2007;60(6):1050–1058.
13. Buzzini SR, Guskiewicz KM. Sport-related concussion in the young athlete. *Curr Opin Pediatr*. 2006;18(4):376–382.
14. Guskiewicz KM, Weaver NL, Padua DA, Garrett WE, Jr. Epidemiology of concussion in collegiate and high school football players. *Am J Sports Med*. 2000;28(5):643–650.
15. Field M, Collins MW, Lovell MR, Maroon J. Does age play a role in recovery from sports-related concussion? A comparison of high school and collegiate athletes. *J Pediatr*. 2003;142(5):546–553.
16. Collie A, Maruff P, Darby DG, McStephen M. The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test-retest intervals. *J Int Neuropsychol Soc*. 2003;9(3):419–428.
17. Rosenbaum AM, Arnett PA, Bailey CM, Echemendia RJ. Neuropsychological assessment of sports-related concussion: measuring clinically significant change. In: Slobounov S, Sebastianelli W, eds. *Foundation of Sports-Related Brain Injuries*. New York, NY: Springer; 2006:137–169.
18. Iverson GL, Lovell MR, Collins MW. Interpreting change on ImPACT following sport concussion. *Clin Neuropsychol*. 2003;17(4):460–467.
19. Duff K, Beglinger LJ, Schultz SK, et al. Practice effects in the prediction of long-term cognitive outcome in three patient samples: a novel prognostic index. *Arch Clin Neuropsychol*. 2007;22(1):15–24.
20. Frazen MD. *Reliability and Validity in Neuropsychological Assessment*. New York, NY: Plenum; 1989.
21. Mitrushina MN, Boone KB, D'Elia L. *Handbook of Normative Data for Neuropsychological Assessment*. New York, NY: Oxford University Press; 1999.
22. Broglio SP, Ferrara MS, Macciocchi SN, Baumgartner TA, Elliott R. Test-retest reliability of computerized concussion assessment programs. *J Athl Train*. 2007;42(4):509–514.
23. Iverson GL, Lovell MR, Collins MW. Validity of ImPACT for measuring processing speed following sports-related concussion. *J Clin Exp Neuropsychol*. 2005;27(6):683–689.
24. Schatz P. Long-term test-retest reliability of baseline cognitive assessments using ImPACT. *Am J Sports Med*. 2010;38(1):47–53.
25. Brandt J, Benedict RHB. *Hopkins Verbal Learning Test-Revised: Professional Manual*. Odessa, FL: Psychological Assessment Resources; 2001.
26. Woods SP, Scott JC, Conover E, et al. Test-retest reliability of component process variables within the Hopkins Verbal Learning Test-Revised. *Assessment*. 2005;12(1):96–100.
27. Benedict RHB. *Brief Visuospatial Memory Test-Revised: Professional Manual*. Odessa, FL: Psychological Assessment Resources; 1997.
28. Benedict RHB. Effects of using same-versus alternate-form memory tests during short-interval repeated assessments in multiple sclerosis. *J Int Neuropsychol Soc*. 2005;11(6):727–736.
29. Reitan RM. *Trail Making Test: Manual for Administration and Scoring*. Tucson, AZ: Reitan Neuropsychology Laboratory; 1992.
30. Kortte KB, Horner MD, Windham WK. The Trail Making Test, Part B: cognitive flexibility or ability to maintain set? *Appl Neuropsychol*. 2002;9(2):106–109.
31. Giovagnoli AR, Del Pesce M, Mascheroni S, Simoncelli M, Laiacona M, Capitani E. Trail Making Test: normative values from 287 normal adult controls. *Ital J Neurol Sci*. 1996;17(4):305–309.
32. Valovich McLeod TC, Barr WB, McCrea M, Guskiewicz KM. Psychometric and measurement properties of concussion assessment tools in youth sports. *J Athl Train*. 2006;41(4):399–408.
33. Smith A. *Symbol Digit Modalities Test Manual*. Los Angeles, CA: Western Psychological Services; 1972.
34. Benedict RH, Duquin JA, Jurgensen S, et al. Repeated assessment of neuropsychological deficits in multiple sclerosis using the Symbol Digit Modalities Test and the MS Neuropsychological Screening Questionnaire. *Mult Scler*. 2008;14(7):940–946.
35. Trenerry MR, DeBoe J, Leber WR. *Stroop Neuropsychological Screening Test: Manual*. Odessa, FL: Psychological Assessment Resources; 1989.
36. Lemay S, Bedard MA, Rouleau I, Tremblay PL. Practice effect and test-retest reliability of attentional and executive tests in middle-aged to elderly subjects. *Clin Neuropsychol*. 2004;18(2):284–302.

37. Franzen MD, Tishelman AC, Sharp BH, Friedman AG. An investigation of the test-retest reliability of the Stroop Color-Word Test across two intervals. *Arch Clin Neuropsychol*. 1987;2(3):265–272.
38. Iverson GL, Lovell MR, Collins MW. Immediate Post-Concussion Assessment and Cognitive Test (ImPACT): normative data. Version 2.0. 2003. http://www.impacttest.com/ArticlesPage_images/Articles_Docs/7ImPACTNormativeDataversion%202.pdf. Accessed May 2, 2007.
39. Hunt TN, Ferrara MS. Age-related differences in neuropsychological testing among high school athletes. *J Athl Train*. 2009;44(4):405–409.
40. Salthouse TA. Decomposing age correlations on neuropsychological and cognitive variables. *J Int Neuropsychol Soc*. 2009;15(5):650–661.
41. Salthouse TA. When does age-related cognitive decline begin? *Neurobiol Aging*. 2009;30(4):507–514.
42. Duff K, Westervelt HJ, McCaffrey RJ, Haase RF. Practice effects, test-retest stability, and dual baseline assessments with the California Verbal Learning Test in an HIV sample. *Arch Clin Neuropsychol*. 2001;16(5):461–476.
43. Bleiberg J, Cernich AN, Cameron K, et al. Duration of cognitive impairment after sports concussion. *Neurosurgery*. 2004;54(5):1073–1080.
44. McCaffrey RJ, Ortega A, Orsillo SM, Nelles WB, Haase RF. Practice effects in repeated neuropsychological assessments. *Clin Neuropsychol*. 1992;6(1):32–42.
45. Segalowitz SJ, Mahaney P, Santesso DL, MacGregor L, Dywan J, Willer B. Retest reliability in adolescents of a computerized neuropsychological battery used to assess recovery from concussion. *Neuro Rehabil*. 2007;22(3):243–251.
46. Schatz P. Long-term test-retest reliability of baseline cognitive assessments using ImPACT. *Am J Sports Med*. 2010;38(1):47–53.
47. Parsons TD, Notebaert AJ, Shields EW, Guskiewicz KM. Application of reliable change indices to computerized neuropsychological measures of concussion. *Int J Neurosci*. 2009;119(4):492–507.
48. Hinton-Bayre AD, Geffen GM, Geffen LB, McFarland KA, Friis P. Concussion in contact sports: reliable change indices of impairment and recovery. *J Clin Exp Neuropsychol*. 1999;21(1):70–86.

Address correspondence to Johna K. Register-Mihalik, PhD, ATC, WakeMed Health and Hospitals, 3024 New Bern Avenue, Raleigh, NC 27610. Address e-mail to jmihalik@wakemed.org.